Research on the method of constructing knowledge graph for e-commerce agricultural products

Hongyi Chi[1,a], Junhui Wu[2]*, Jianglong Liu[1,b], Zixi Chen[1,c], Jue Gong[1,d]

[a]2460327840@qq.com; [*]junhui_wu@163.com; [b]2233022@tongji.edu.cn
[c]2897231674@qq.com; [d]874321116@qq.com

[1]School of Electronic and Information Engineering Tongji University, Shanghai, China
[2]National Engineering Research Center of Protected Agriculture Tongji University, Shanghai, China

Abstract. The traditional e-commerce agricultural products information service has problems such as insufficient information mining and correlation, and inability to meet users' deep needs. From the user's information needs, this paper proposes the application of knowledge mapping technology to extract and fuse unstructured data, to achieve deep mining and association of heterogeneous e-commerce agricultural products and user comment data from multiple sources, to solve the problem of poor information between supply and demand, and at the same time, to take tea as an example of designing the construction method of e-commerce agricultural products knowledge mapping.

Keywords: knowledge graph; agricultural e-commerce

1. Introduction

In recent years the Internet, big data and other information technology in the field of agricultural products e-commerce is widely used, information technology and its application to solve the problem of poor information on agricultural products business provides a new direction. However, the current traditional e-commerce agricultural products information service system has problems such as insufficient information mining and correlation, single service content, and inability to meet the deep-seated needs of users.[6] Therefore, how to effectively organize and merge a large amount of unstructured agricultural product attribute and comment information, and carry out in-depth mining and organization of information, so that the orderly and efficient circulation of agricultural products has become a difficult problem for the development of agricultural product e-commerce.

In this paper, firstly from the user's information needs, based on the organization and mining of agricultural product information in the e-commerce environment, to provide knowledge[7] support for the e-commerce agricultural product information service system. Secondly, the labelling model that effectively reacts to the overall characteristics of agricultural products by mining and analyzing product information data[1]. Finally, using knowledge mapping technology to organize, store and manage unstructured data, to achieve the depth of mining and

ECIT 2024, March 15-17, Changsha, People's Republic of China

DOI 10.4108/eai.15-3-2024.2346179

correlation of multi-source heterogeneous e-commerce agricultural products and user state review data, to enrich the support of consumer purchasing decisions, mining agricultural production and management of the main body of the demand for data sources, as an example of tea commodities to ultimately design the construction of the e-commerce agricultural products knowledge mapping method.

2. Data set construction and pre-processing

2.1 Data cleaning

For the crawled data related to tea products, the data is cleaned by removing useless fields, content cleaning, and de-duplication processing in three ways to prepare the corpus for subsequent knowledge extraction and so on. As shown in Figure 1. below.

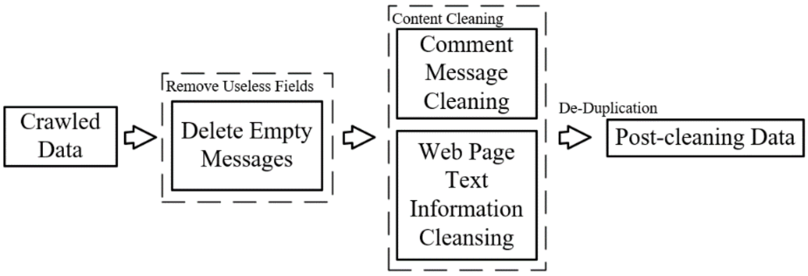


Figure 1. Data Preprocessing

- Empty lines, empty strings and meaningless data: Filtering operations are performed on such data to remove empty information and improve the semantic information content of the collected data.

- Content cleansing: It is divided into comment information cleaning and page text information cleaning. Comment information cleaning needs to filter the spam comments. In this paper, the word count of single comment information is counted, and the comment information whose word length is less than 5 is regarded as invalid comment, so this part of comment is deleted; and the single repeated comment sentence is deleted. Also with the help of regular expressions to achieve the filtering and cleaning of impurity labels, to ensure the neatness of the text data.

- De-duplication process: there may be a large amount of duplicate data in the crawled information, and the duplicate data will have an impact on the quality of the subsequent mapping data, thus affecting the service efficiency of the e-commerce agricultural product information service system, so it is necessary to reduce the duplicate data.

2.2 Domain dictionary construction

The data in the dictionary is collated from the semi-structured tea product related data in the crawled product detail page, and its label content can be directly parsed using a parser. The structured data can be directly corresponded to the corresponding attributes when entering the database, without knowledge extraction. Take the tea brand dictionary as an example, the crawled brand data is manually filtered and organized into the "top ten best-selling brands of tea in 2021" as shown in Table1. below, and the schema layer of this paper will be developed around these ten brand entities, gradually filling in the corresponding tea commodity entities, farmers and their relationships with each other.

Table 1. Top 10 tea brands

Serial Number	Brand Name
1	China tea
2	WUYUTAI TEA
3	ZHANGYIYUAN TEA
4	JING HUA TEA
5	FENGPAI TEA
6	SANWANCHANG TEA
7	TIANMU YUNLU Tea
8	Zhejiang Tea Group
9	LUZHENGHAO TEA
10	West lake tea

2.3 Dataset labelling

In this paper, the crawled data is divided into clauses, ending with full stops, question marks and exclamation marks, to prepare the data for the subsequent real data for the preparation of experiments for the subsequent joint extraction model of corpora and relations. After processing the data in the above way, 1790 items are manually selected as sentences. After processing the data in the above way, 1790 items were manually selected and manually labelled in the form of "BIO" to be used in the subsequent extraction experiments. The data were manually selected and labelled as "BIO" for subsequent extraction experiments.

3. Knowledge graph construction

3.1 Constructing knowledge graph schema layers

The construction of the tea knowledge graph pattern layer is a combination of the concept of product image and the existing tea classification catalogue in the e-commerce platform. In this paper, we refer to the current multi-level classification navigation in several major e-commerce platforms that are divided according to the characteristics of tea brands, types, etc., and collate the concepts, entities and their hierarchical relationships in the product image schema layer of the tea category specifically as shown in Figure 2.

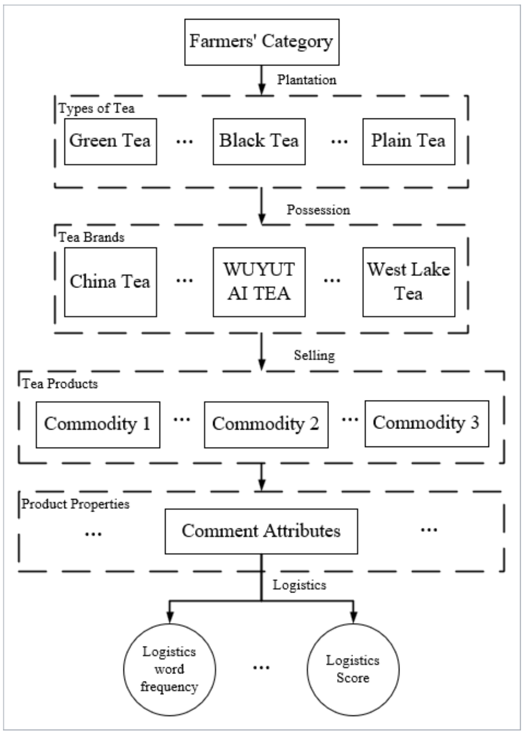


Figure 2. Pattern layer of the commodity spectrum map for the tea category

Knowledge graph is a large-scale knowledge network that connects entities and relationships with each other. In order to associate commodity entities with each other and agricultural production and management subjects to better understand the attributes of their products, four types of entities, namely, "brand type", "farmer type", "tea type", and commodity review type, are introduced to associate with commodity entities.

3.2 Knowledge graph data layer construction

The extraction of tea trade names is based on the rule and word list matching extraction method. By analyzing the compositional characteristics of the named entities of tea on each e-commerce platform, it is found that the trade names are generally composed of the "brand name + origin + commodity grade + tea type + packaging form" of tea and the weight of the product "g". Therefore, the extraction of tea trade names firstly obtains the noun entity through word list matching, then obtains the product quality through rule extraction, and finally combines the results of word splitting into trade name results.[8]

Experimental Procedures:

- Dictionary Construction of Trade Names: In this paper, based on the crawled semi-structured related attribute-valued nouns, we integrate them after preprocessing to construct a noun dictionary of trade name composition stored in txt. Part of the noun dictionary is shown in Table 2:

Table 2. Partial example of a word list of trade name components

Lexical Epithet	Match
Brand Name	Feng Brand, Jinghua, Lu Zhenghao, Sandwiching, Tianmu Yunlu, West Lack Brand, Zhang Yi yuan, Tian Xiang Brand, Chifeng, China Tea
Place of Origin	Yunnan, Fenagling, Yunnan, West Lake, Chifeng Mountain, Dongxiang, Fuding, Suzhou, Donating, Suzhou Donating Ting, Fujian, Wuyi Shan, Wuyi Shan
Product Grade	Premium, Level 1, Level 2, Level 3
Types of Tea	Black Tea, High Mountain Black Tea, Dian Hong Black Tea, Purer Tea, Jasmine Tea, Oolong Tea, High Mountain Green Tea, Green Tea, Longjing Tea, Balouch Tea, Jejune Tea
Packaging Form	Cans, Gift Boxes, Double Can Combos, Bags, Boxes, Traditional Paper Bags, Bulk, Gift Boxes

- Trade Name Extraction: Based on the above characteristics of the tea name entity[4] the tea name entity extraction rules are summarized for the tea name entity, and the specific algorithm steps are described as shown in Table 3.

Table 3. Trade name entity extraction algorithm

Algorithm 1 Trade Name Entity Extraction Algorithm

Import: Tea trade name composition word list set, crawled tea trade name set List1.

Export: Standardized set of tea trade names List2.

Step1: Select the element L1 in List1 to match the noun entity according to the lexicon S1, if there exists a successful matching noun entity, it will be stored in List2. the other elements in List1 match the lexicon S1 in turn;

Step2: The other elements in List1 match the word list S2...S5 in turn, and the noun entities that match successfully are stored in List2. S5, the successful noun entities will be stored in List2;

Step3: Select element L1 in List1, match according to the rule, if "g" appears in L1 and a number appears before "g", connect the number and "g" and extract it as a named entity of the product and store it in List2. If "g" appears in L1 and a number appears before "g", then connect the number and "g" and extract it as a product named entity, and store it in List2. other elements in List1 are matched according to the rule in turn, and store them in List2;

- Sampling results and analyses: According to the above experimental steps, the 350 tea trade name entities crawled were extracted.

The data sources of this paper are tea product information of e-commerce companies such as Jindong and Taobao, tea entry information of Wikipedia and tea related webpage text.[3] The crawled corpus is manually annotated and proofread in the unit of sentence to get the e-commerce tea related text dataset for experimental use. Randomly selected 20% as the test set, and randomly selected 10% as the validation set in the training set. The statistics related to the dataset after manual annotation are shown in Table 4.

Table 4. Statistical data relating to data sets

	Training Set	Test Set	Subtotal
Labels	1432	358	1790

In this paper, the model results are evaluated using three measured metrics: check accuracy, recall and F1 value[2], which are calculated as shown in equations (1), (2), and (3):

$$P = \frac{T_P}{T_P+F_P} \times 100\% \quad (1)$$

$$R = \frac{T_P}{T_P+F_n} \times 100\% \quad (2)$$

$$P = \frac{2PR}{\mathrm{P+R}} \times 100\% \quad (3)$$

(1) where: *Tp* is the number of correctly judged by the model and *Fp* is the number of irrelevant ones identified by the model. Eq. (2) In Eq. (2): *Fn* is the number of the ones that are not identified by the model. For the extraction model algorithm in this paper, the correct extraction result can be obtained only when the entity boundary, relationship category and entity label of each entity in the relationship pair are identified correctly.

In order to prove that BERT is more advantageous than the commonly used word vector representation, BIGRU-CRF sequence annotation model is used for comparison experiments. In this paper, the BERT-BIGRU-CRF sequence annotation model is trained at Epoch 18, and the performance of the validation set data is no longer significantly improved, and some of the main hyperparameters selected are shown in Table 5.

Table 5. Parameters of the BERT-BIGRU-CRF sequence annotation model

Parametric	Retrieve a Value
Learn_rate	5e-5
clip	0.5
Max_seq_length	128
Batch_size	16
Dropout_rate	0.3
dimension of BIGRU	256

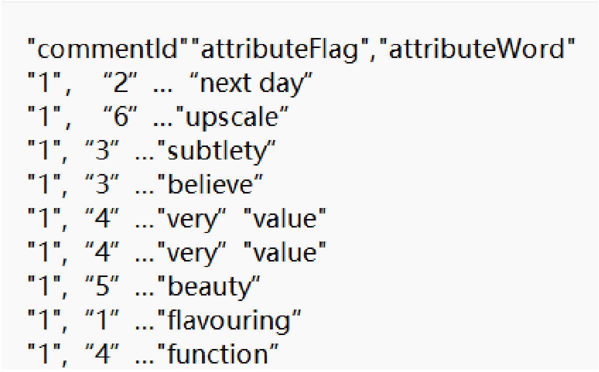
Results and analyses: The joint entity-relationship extraction model proposed in this paper for tea domain is tested using a test set, and the results of relationship entity extraction are shown in Table 6. The BERT-BIGRU-CRF model in this paper improves the F1 value by 6.1%, so BERT is more advantageous than the commonly used word vector representation.

Table 6. Table of model test results

Modelling	P	R	F1
BIGRU-CRF	0.559	0.464	0.507
BERT-BIGRU-CRF	0.620	0.633	0.627

Comment attribute word frequency and sentiment word acquisition experiment data from crawling to the head of the Jindong, Taobao and other tea e-commerce sites, after the data pre-processing of the collected comment data, manually selected 350 of the tea commodity data, to obtain its first 200 comment information, a total of 7,000 pieces of tea comment data used for this paper's sentiment analysis extraction.

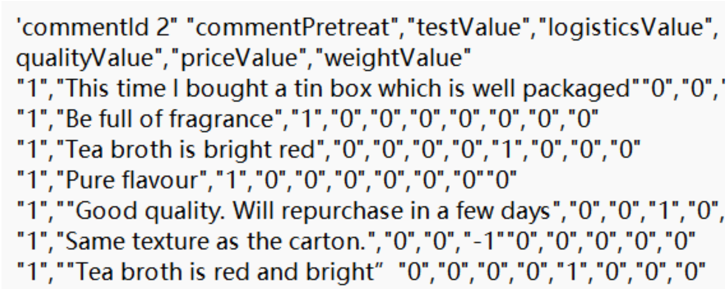
Experimental procedure: In this experiment, the processed tea review information is batch imported into the fine-grained sentiment analysis method of agricultural product reviews developed by this group in list mode, and the tea product review information is analyzed in eight aspects, namely, taste, logistics, quality, price, portion, color, packaging, and after-sale service, and the sentiment intensity scores of different review attribute levels and different attribute word frequencies of each review information are returned, and the returned part of the results are shown in Fig. 3



```
"commentId""attributeFlag","attributeWord"
"1",    “2”  ...  “next day”
"1",    “6”  ..."upscale”
"1",  “3”  ..."subtlety”
"1",  “3”  ..."believe”
"1",  “4”  ..."very”  "value"
"1",  “4”  ..."very”  "value"
"1",  “5”  ..."beauty”
"1",  “1”  ..."flavouring”
"1",  “4”  ..."function”
```

Figure 3. Attribute word frequency results

Among them, commentId represents the commodity number analyzed, attributeFlag represents the fine-grained category labels of the commodity's review attributes (1: taste; 2: logistics; 3: quality; 4: price; 5: portion; 6: color; 7: packaging; 8: service), attributeWord represents the frequency of the relevant attribute words extracted, and after statistical processing, we obtain the top 5 word frequencies of attribute class of each commodity, and organize them into the structure of "Attribute Class - Attribute Word Frequency" to put them into the atlas as a display of the attribute word frequencies of the commodities. As shown in Fig. 4.



```
'commentId 2" "commentPretreat","testValue","logisticsValue",
qualityValue","priceValue","weightValue"
"1","This time I bought a tin box which is well packaged""0","0",'
"1","Be full of fragrance","1","0","0","0","0","0","0","0"
"1","Tea broth is bright red","0","0","0","0","1","0","0","0"
"1","Pure flavour","1","0","0","0","0","0","0""0"
"1",""Good quality. Will repurchase in a few days","0","0","1","0",
"1","Same texture as the carton.","0","0","-1""0","0","0","0","0"
"1",""Tea broth is red and bright”  "0","0","0","0","1","0","0","0"
```

Figure 4. Results of comment attribute-level sentiment intensity scores

Where commentId_2 represents the product number to be analyzed and comment Pretreat represents the comment to be analyzed; the 8 columns of the comment scoring are taste, logistics, quality, price, portion, color, packaging and service in that order. In the sentiment label table, -1 represents negative sentiment, 1 represents positive sentiment and 0 indicates that the comment does not mention the attribute. In the emotional intensity table, less than 0 represents negative emotion, more than 0 represents positive emotion, and 0 indicates that the attribute is not mentioned in the comments.

The output results of the reviews need to be quantified in terms of the emotional strength score of the product attributes. Through the interface returns a single commodity "attribute-attribute value", therefore, it is necessary to process the returned results into the form of "attribute class-score" in order to represent the results of the review of a certain attribute of this commodity, so as to achieve the result of the correspondence between the tea commodity and the emotional intensity score of a certain attribute. Therefore, it is necessary to transform the returned result into the form of "attribute class-score" to represent the comment result of a certain attribute of this product, so as to achieve the correspondence between tea products and the emotional intensity score of a certain attribute. For a certain attribute of a product, after removing the result with a score of 0, the remaining m relevant comment scores, and after adding up the m scores, the result is n, then the sentiment score of the attribute class is n/m. The structure of the obtained "attribute class-score", i.e., the evaluation of a certain attribute of a product by all the users of the feedback, serves as the entity knowledge of the mapping comment analysis.

Experimental results: The results of the analysis of the reviews of a particular tea commodity, compiled after the above processing and analysis, are shown in Table 7.

Table 7. Results of Analyzing Tea Commodity Reviews

Trade name	Subject of evaluation	Evaluation of word frequency	Combined score (-1, 1)
Trade A	wrap	Product packaging, jars, gift boxes, etc.	0.8
	component	Size, Gift, Scale	0.7
	prices	Markdowns, discounts, events	0.5
	texture	Taste, flavor, etc.	-0.3
	quality	Product quality, materials, ingredients, etc.	-0.4
	color	Shades, colors, shellability	-0.1
	aftermarket	Customer Service, Follow-up, Acceptance	-0.9
	logistics	Courier, speed, delivery	0.5

In this paper, the acquired words of tea-related entities and attributes from different sources are manually arranged to construct a word list of data to be fused.[5] In this paper, we use the open source Python proxemics toolkit genism module for model parameter setting and process training to judge the similarity of different source information. The training parameters of the word2vec algorithm model using genism are shown in Table 8 below, and the word vector model "Chinese_tea. model" is obtained after training, and the specific steps of the algorithm are shown in Table 9.

Table 8. Word2vec model training parameter settings

Parametric	Value	Account
Sg	0	0 means using the CBOW model
size	200	Dimension of the word vector obtained from training
window	10	Window size for training
Min_count	5	Ignore words with a frequency lower than the specified min_count.
hs	1	Using the Hierarchical SoftMax method

Table 9. Flow of word2vec-based synonym alignment algorithm

```
Algorithm 2 Synonym alignment algorithm based on word2vec
Import: model Chinese_tea.model , word list List1={Entity1{e1,e2,e3...} ,Entity2{e1,e2,e3,...}...Entity{e1,e2,e3,...}}, threshold M=0.7
Export: List2={E1,E2,E3…En}

Step1:Take a collection of similar entities from List1 Entity1{e1,e2,...}
Step2:Iterate through the set and calculate the similarity between e1 and the entities in Entity1, if the similarity > M, then determine that e1 is synonymous with the other entities and align the two terms.
Step3:Execute Step2 until List1 is completely traversed.
Step4:End of algorithm
```

3.3 Knowledge graph construction

After the acquisition and extraction of heterogeneous data from multiple sources as described above, the structured data is organized into the form of ternary groups, and the data nodes are imported into the Neo4j graph database in batch through the creation of the Cypher statement Load csv to create the relationship between entities and complete the drawing of the graph. [9]After the mapping is completed, there is a query search box in the database interface, you can use the Cypher statement for the visual query of the knowledge graph, the interface can return the results of the query, the results can be returned to the node graph or table, this paper only selects some of the tea class, commodity class (including attribute entities and comments on the entity), farmers, brand entities for the visualization of the display. You can see the results shown in Figure 5.

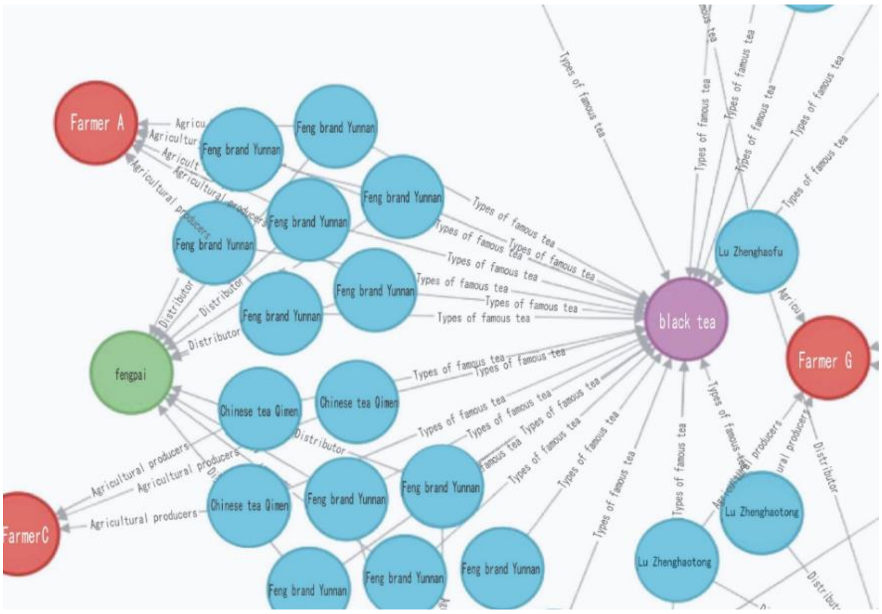


Figure 5. Knowledge spectrum of tea products

4. Summary and future outlook

As the attributes related to agricultural products on e-commerce platforms and user feedback information are scattered in multiple data sources on the network, it greatly affects consumers and agricultural production and management subjects to find their own information of concern and influences the basis for decision-making.[10] Therefore, from the perspective of user needs, this paper proposes the application of knowledge mapping technology to extract and fuse unstructured data to achieve deep mining and correlation of heterogeneous e-commerce agricultural products and user platform data from multiple sources. On the basis of the attribute information and comment information of agricultural products, the nodes of brands, farmers and types of agricultural products are introduced to construct the knowledge graph of e-commerce agricultural products, so that the different agricultural entities in the e-commerce platform are also associated with each other to promote the further development of e-commerce sales of agricultural products. The product knowledge graph constructed in this paper contains static information and dynamic comments on agricultural products, and is currently combined with user access data logs to update and maintain the graph data on a fixed cycle. In the era of information explosion, fixed-cycle maintenance may lead to untimely updating of information and information lag problem, which in turn affects the query results of the system. Therefore, the implementation of dynamic updating of the atlas should be considered in subsequent research.

References

[1] Wang Ying. Research on product portrait construction based on knowledge graph [D]. Nanjing University of Science and Technology, 2018.

[2] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation[C]. European conference on information retrieval, 2005: 345-359.

[3] Kaiying D, Senpeng C, Jingwei D. On optimisation of web crawler system on Scrapy framework[J]. International Journal of Wireless and Mobile Computing, 2020, 18(4): 332-338.

[4] Zhang X, Li C, Du H. Named Entity Recognition for Terahertz Domain Knowledge Graph based on Albert-BILSTM-CRF[C]. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020.

[5] R.J. Chen, S.Y. Zheng, Y.X. Zhu. Joint extraction of entity relations by fusing entity category information[J]. Computer Engineering, 2022, 48(03): 46-53.

[6] Kauffman R J, Li T, Van Heck E. Business network-based value creation in electronic commerce[J]. International Journal of Electronic Commerce, 2020, 15(1): 113-144.

[7] Zhang Jixiang, Zhang Xiangsen, Wu Changxu, et al. A review of knowledge graph construction techniques[J]. Computer Engineering, 2022, 48(03): 23- 37.

[8] Nickel, Maximilian, Murphy, et al. A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction[J], 2019.

[9] Gomez-Perez J M, Pan J Z, Vetere G, et al.: Enterprise knowledge graph: An introduction, Exploiting linked data and knowledge graphs in large organisations: Springer, 2019: 1-14.

[10] Yan Kai. Research on consumer demand of cultural and creative agricultural products under e-commerce sales channel[J]. Logistics Science and Technology, 2020, 43(06): 52-53.